Published in final edited form as:
J Am Coll Radiol. 2012 November ; 9(11): 788–794. doi:10.1016/j.jacr.2012.05.020.

Radiologist Agreement for Mammographic Recall by Case Difficulty and Finding Type

Tracy Onega, PhD[1], **Megan Smith**[2], **Diana L. Miglioretti, PhD**[2,*], **Patricia A. Carney, PhD**[3], **Berta Geller, EdD**[4], **Karla Kerlikowske, MD**[5], **Diana SM Buist, PhD**[2], **Robert D. Rosenberg, MD**[6], **Robert Smith, PhD**[7], **Edward A. Sickles, MD**[5], **Sebastien Haneuse, PhD**[8], **Melissa L. Anderson**[2], and **Bonnie Yankaskas, PhD**[9]

[1]Department of Community & Family Medicine, Dartmouth Medical School, Hanover, NH

[2]Group Health Research Institute, Seattle, WA

[*]Department of Biostatistics, University of Washington, Seattle, WA

[3]Departments of Family Medicine and Public Health & Preventive Medicine, Oregon Health & Science University, Portland, OR

[4]University of Vermont, Burlington, VT

[5]Departments of Medicine and Epidemiology/Biostatistics and Radiology, University of California San Francisco, CA

[6]Department of Radiology, University of New Mexico, Albuquerque, NM

[7]American Cancer Society, Atlanta, GA

[8]Department of Biostatistics, Harvard School of Public Health, Boston, MA

[9]Department of Radiology, University of North Carolina-Chapel Hill, NC

Abstract

INTRODUCTIONS—To assess agreement of mammography interpretations by community radiologists with consensus interpretations of an expert radiology panel, to inform approaches that improve mammography performance.

METHODS—From six mammography registries, 119 community-based radiologists were recruited to assess one of four randomly assigned test sets of 109 screening mammograms with comparison studies for no recall or recall, giving the most significant finding type [mass, calcifications, asymmetric density or architectural distortion] and location. The mean proportion of agreement with an expert radiology panel was calculated by cancer status, finding type, and difficulty level of identifying the finding at the woman, breast, and lesion level. We also examined concordance in finding type between study radiologists and the expert panel. For each finding type, we determined the proportion of unnecessary recalls, defined as study radiologist recalls that were not expert panel recalls.

RESULTS—Recall agreement was 100% for masses and for exams with obvious findings in both cancer and non-cancer cases. Among cancer cases, recall agreement was lower for lesions that

were subtle (50%) or asymmetric (60%). Subtle non-cancer findings and benign calcifications showed 33% agreement for recall. Agreement for finding responsible for recall was low, especially for architectural distortions (43%) and asymmetric densities (40%). Most unnecessary recalls (51%) were asymmetric densities.

CONCLUSION—Agreement in mammography interpretation was low for asymmetric densities and architectural distortions. Training focused on these interpretations could improve mammography accuracy and reduce unnecessary recalls.

Keywords

mammography; breast cancer; screening; agreement

INTRODUCTION

The effectiveness of screening mammography depends on the accuracy of image interpretation. However, the interpretive performance of US radiologists in community practice varies widely for screening mammography [1]. Notable interobserver variability is seen, even when interpreting the same mammographic images [2-7]. Most studies comparing radiologists' interpretations of screening mammogram test sets have focused on agreement of the Breast Imaging and Reporting Data System [BI-RADS®] assessment [4, 6, 7-9]. Berg et al. [5] found considerable variability in the assessment of 103 screening mammograms interpreted by five experienced radiologists. Ciatto et al. [6] and Kerlikowske et al. [4] reported moderate interobserver agreement in studies of 100 examinations interpreted by 12 radiologists, and 2616 mammograms interpreted by two radiologists, respectively. Elmore et al. [3], Lazarus et al. [9], Venta et al, [10] and Skaane, et al. [11] reported substantial interobserver variation in studies of 50 to 1147 mammograms, with 5 to 10 interpreting radiologists.

Agreement of mammography interpretations by finding type (calcification, mass, asymmetry or architectural distortion) and case assessment difficulty has not been well characterized. Berg et al. examined agreement of five radiologists experienced in mammography for lesion classification and defining features and found high interobserver agreement (Kappa=0.75) for lesion classification, but not for features and management [5]. However, this prior study was limited by small sample size and lack of an expert consensus standard. Determining which mammographic findings have poor agreement with a well-defined standard could identify areas of focus for interventions aimed at improving interpretive accuracy.

The purpose of this study was to assess agreement between mammography interpretations of community radiologists and a gold standard, defined as consensus interpretations of an expert panel. We hypothesized that certain findings or features would be mainly responsible for lack of agreement. These would be potential areas to target for improving mammography performance. Our study includes the largest number of participating radiologists to date for a mammography interpretation agreement study. We analyzed variability in interpretation of screening mammograms by finding and by difficulty of identification, examining agreement of 119 community radiologists to a consensus-based standard developed by three mammography experts.

METHODS

Study Population

This study was conducted within six breast screening registries of the National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC): Carolina Mammography

Registry, Group Health Surveillance Project in Washington State, New Hampshire Mammography Network, San Francisco Mammography Registry, New Mexico Mammography Project, and Vermont Breast Cancer Surveillance System [13, 14]. Radiologists interpreting mammography at BCSC facilities between January 2005 and December 2006 were invited to participate. In addition, we invited 103 radiologists from outside the BCSC in Oregon, Washington, North Carolina, San Francisco, and New Mexico. Of the 469 invited radiologists, 148 (31.6%) provided informed consent and 119 (80.4%) of these completed the study. Radiologists received up to eight Category I continuing medical education credits for interpreting a test set. Each site received Institutional Review Board approval for study activities. Active or passive consent and/or waivers are obtained at each site from women receiving mammograms. Identities of women, physicians, and facilities are protected by a Federal Certificate of Confidentiality and other protections. All procedures comply with the Health Insurance Portability and Accountability Act.

Test set development

Test set development is described in detail elsewhere [15]. Briefly, we developed four test sets differing by cancer prevalence and case difficulty of 109 mammograms each, sampled from 130 screening mammograms interpreted by BCSC radiologists from 2000 to 2003. Some of the 130 cases were included in one or more test sets based on the desired composition of each test set. For example, test sets 1 and 2 contained 13% subtle cases, while test sets 3 and 4 contained 30% subtle. In order to achieve the needed proportions, some of the subtle cases were used in two or more of the test sets. We used screening examinations for women aged 40-69 years with a previous mammogram within the prior 11-30 months that could be used for comparison views. We excluded women with a history of breast cancer or breast augmentation. Each test set case consisted of craniocaudal (CC) and mediolateral oblique (MLO) views of each breast (4 views per woman for each screening and comparison examination).

Custom-designed software for viewing images and collecting interpretations was created in collaboration with the American College of Radiology (ACR), whose staff professionally digitized and uploaded the films into the software on DVDs.

Three expert radiologists participated in test set development, all of whom are senior radiologists in academic medical centers who teach and specialize in breast imaging, and have all been past presidents of the Society for Breast Imaging (SBI) and/or recipients of SBI's Gold Medal award. The experts reviewed 320 digitized screening studies and exclude those with poor image quality or marks that were not removed when digitized, resulting in the 130 studies of the test set. Before administering the test, each digitized mammogram was independently reviewed by the three expert radiologists using the test set software. Each expert made an independent assessment, indicating whether the patient should be recalled for a finding visible on the digitized image. Each expert identified the most significant finding and classified that finding as a mass, calcification, asymmetric density, or architectural distortion and assigned a level of difficulty of finding identification (obvious, intermediate, subtle). Consensus expert opinion was taken to be the agreement of at least two of three experts for each measure. A consensus meeting resolved 68 cases for which all three expert radiologists disagreed. All finding types and descriptors used in this study were based on this expert consensus. For each finding recalled by the panel of exerts, one expert defined the lesion location ("region of interest") by drawing a rectangle around it.

Test Set Administration

Consenting radiologists were randomized to one of the four test sets using a block randomization scheme with stratification by BCSC registry (or site for non-BCSC

radiologists) and by whether the radiologist had interpreted at least 30 breast cancer cases in the BCSC database. This ensured sufficient power to examine clinical performance measures related to test set performance as part of the larger study.

Test set DVDs were sent to each consenting radiologist with their test set assignment and an instruction sheet for software system registration. Radiologists used a computer of their choice, or a laptop provided by the study, with a DVD reader and display requirements that allowed viewing two images concurrently, with sufficient resolution.

Radiologists were instructed to interpret the images as they would in clinical practice, except that they were requested to record only the most significant finding, if any. Radiologists were informed that the test sets had been cancer-enriched but the specific prevalence of cancer cases was not revealed. For each case, radiologists indicated whether they would recall the case using the software, which defined recall/no recall based on the American College of Radiology BI-RADS lexicon definition of recall = codes 0, 4 and 5; and no recall= codes 1 and 2. For recall cases, radiologists selected a finding type (mass, calcification, asymmetric density, or architectural distortion), and indicated the location by clicking on the image(s), which translated the location into coordinates stored by the software.

Outcome Measures

Our main outcome measure was case assessment (recall/no recall). Results were stratified by cancer status of cases. Cancer cases had either ductal carcinoma *in situ* or invasive breast cancer diagnosed within one year of the mammogram. Non-cancer cases did not have a breast cancer diagnosis within 24 months after the mammogram. We defined the recall of a non-cancer case as *appropriate* if the expert panel determined additional imaging was necessary because the chance of cancer based on the screening results was sufficiently high. *Unnecessary recalls* were cases that did not require additional work-up according to the expert panel.

Analyses

Analyses of recall agreement were conducted at the woman, breast, and lesion level to help understand the level at which interpretive agreement is most problematic and may have the most clinical impact and to refine the agreement to test whether it is important to tie agreement to the actual lesion vs. identifying the correct breast. A study radiologist was considered as recalling the same breast as the experts if the finding location indicated by the radiologist was in the same breast as the region of interest indicated by the experts. A study radiologist was considered to recall the same significant breast lesion as the experts if the finding location indicated by the radiologist was within the gold standard region of interest. We calculated the proportion of mammograms recalled by study radiologists separately by cancer status, the difficulty level of identifying a finding, and lesion type. More specifically, we calculated the percent recall (agreement with expert consensus) for each study radiologist, separately by cancer status and case characteristics (difficulty and finding type as defined by the expert panel), and report the median and interquartile range (IQR) of study radiologist percent agreement with the gold standard. We graphed the proportion of cases recalled by 0-25%, 26-50%, 51-75%, and 76-100% of study radiologists, separately by case difficulty and finding type for cancer cases and appropriate recalls.

We examined whether consistent nomenclature was used to classify abnormal findings, or whether there was variability in the finding type classification when the same lesion was recalled. To do this, we report the distribution of finding type classifications given by the study radiologists when they recalled the same lesion as the expert panel, stratified by the

gold-standard finding of the expert panel and by level of difficulty. We calculated the distribution of finding type for all cases recalled by the community radiologists in relation to whether the experts recalled it.

RESULTS

The majority of participating radiologists (64%) reported interpreting mammography for more than 10 years (Table 1). Few had breast fellowship training (13%) or self-identified as a breast specialist (8%), and most were not affiliated with an academic institute (86%). Over half (51%) of participating radiologists reported working three or more days a week in breast imaging and 63% reported interpreting an average of ≥1000 mammograms per year during the previous 5 years.

Variation in recall among community radiologists

The median percent of mammograms recalled by the study radiologists varied by level of difficulty and finding type (Table 2). Overall, a median of 80% (IQR: 70-87%) of cancers were recalled at the woman level, 73% (IQR: 67-83%) at the breast level, and 67% (IQR: 55-73%) at the lesion level (Table 2). At the woman level, recall was much less likely for subtle cancers (50%; IQR: 40-75%) than intermediate (86%; IQR: 69-100%) or obvious (100%; IQR: 83-100%). The median percent of subtle cancers recalled decreased at the breast and lesion levels (Table 2). Among cancers, asymmetric densities (as defined by the experts) had the lowest median percent recall (woman level: 60%; IQR: 50-75%), with 50% recalled at the lesion level (IQR: 40-75%). Masses were the most likely to be recalled, with 100% recalled at the woman (IQR: 83-100%) and breast level (IQR: 78-100%) (Table 2).

Findings for non-cancer cases with lesions that were appropriate recalls were similar to cancer cases: subtle benign lesions were recalled less often than intermediate and obvious lesions (Table 2). Benign calcifications recalled by the experts were “appropriately” recalled only one-third of the time by study radiologists (33% median recall; IQR: 33-67%). The median rate of unnecessary recall among non-cancers was 26% (IQR: 16-33%).

Variation between community radiologists and the expert panel in recall recommendation

We further examined agreement in recall recommendations between the study radiologists and the expert panel. We report the percent of readers who recalled cases, stratified by case characteristics. Overall, an increasing level of difficulty reduced observed agreement in recall, as did measuring agreement at the lesion level for both cancers and non-cancers (Figures 1a and 1c). For non-cancers, agreement between the majority of study radiologists and the expert panel was markedly lower for appropriate recall of intermediate and subtle lesions than for obvious lesions (Figure 1c). For recalling asymmetric densities that were cancers, majority agreement with the expert panel was 44% at the lesion level (Figure 1b). However, at the woman level, the majority of radiologists (51% or higher) agreed with the expert panel for most finding types (mass: 100%, calcification: 83%, asymmetric densities: 77%, and architectural distortion 100%). For recall of non-cancer lesions, the majority of radiologists agreed with the expert panel only for masses (100% for woman and breast level) (Figure 1d). For recall of non-cancer cases at the woman level, majority agreement with the expert panel was relative low for calcifications (33%), asymmetric densities (66%), and architectural distortions (50%) (Figure 1d).

We found an overall high concordance in nomenclature for masses and calcifications (79.9% and 91.7%, respectively), with much lower concordance for asymmetric densities (40.7%) and architectural distortion (45.3%) (Table 3). For subtle cases, lack of concordance for asymmetries and architectural distortion appeared to be caused by classification of

asymmetric densities as distortions and *vice versa* (Table 3a). For intermediate cases, both asymmetric densities and architectural distortions were frequently classified as masses (46.1% and 32.0%, respectively) (Table 3b). Obvious cases showed high concordance for all finding types (Table 3c).

The goal of breast screening effectiveness research is achieving good sensitivity with low false positive rates, for example by discovering ways to minimize unnecessary recalls. Therefore, to determine if certain findings were associated with more unnecessary recalls, we examined appropriate and unnecessary recalls of non-cancers by finding type (Table 4). The most frequent appropriate recalls, defined as agreement with the expert panel, were for masses (50%). Calcifications accounted for 14% of appropriate recalls, and only 7% of unnecessary recalls. Asymmetric densities were 40% of all non-cancer findings, but only 26% of appropriate recalls and 51% of unnecessary recalls. Architectural distortions accounted for only 10% of appropriate recalls and 15% of unnecessary recalls.

DISCUSSION

This test-set study is the largest to date to examine agreement between radiologist recall and a consensus-derived gold standard interpretation. It is the first to examine the factors of interpretive difficulty and finding type at the woman, breast, and lesion level. As expected, agreement was high for obvious findings, and markedly lower for subtle findings, most notably among non-cancer cases. Overall, agreement by finding type was relatively high for cancer cases compared to non-cancers, but calcifications, architectural distortions, and asymmetries all contributed to lower agreement. Much of the difference in agreement appeared to be due to nomenclature, particularly for architectural distortion and asymmetries. Most unnecessary recalls by the participating radiologists were for asymmetries.

Our analysis of woman-, breast-, and lesion-level findings suggested that radiologists may arrive at the same recall decision even if they are not basing their decision on the same lesion. We also found that woman- and breast-level agreement was consistently stronger than lesion-level. The implications of these findings may be a similar clinical course for any cases for which there is at least breast-level, but particularly lesion-level agreement, because in radiologists recommend follow-up that is pursued to a final interpretation. Accuracy at the breast level may indeed lead to appropriate work-up and treatment, however there will be times when the lesion is benign or a cancer lesion is missed, or when a more efficient workup and/or better clinical course would result if the specific lesion is correctly identified on screening.

Nonetheless, specific lesions may require particular courses of clinical care, and differences in finding type may alter management. An advantage of our multilevel approach is that we were able to identify the most challenging finding types, which will inform strategies to improve interpretive performance. By focusing on non-mass findings, particularly in intermediate and subtle cases, responsive educational interventions can be designed to yield clinically important improvements.

This study provides a unique perspective on variability in mammography interpretation by measuring agreement with a well-defined standard interpretation. Several other studies examined interobserver agreement, particularly for assessing performance and use of the BI-RADS lexicon [5-7, 16]. A focus on recall/no recall, as in this study has high clinical relevance as it determines whether a woman will undergo further follow-up, which provides additional information for a final assessment. Previous studies of lesion agreement focused on detailed characterization of mass characteristics [5, 16], and we found the highest level of

agreement between study radiologists and the expert panel for masses. Our results showed the lowest agreement for architectural distortion and asymmetric densities, consistent with studies that suggest that classification of finding type rather than detection reduces interobserver agreement [5, 6, 16].

Several important aspects of this study should be noted in interpreting the results. First, radiologists were in a testing situation rather than a usual clinical care setting. Gur et al. reported significantly lower performance level among radiologists (n=9) in the laboratory compared to the clinic, and lower inter-reader dispersion in a clinical setting [17]. In contrast, another study comparing clinical and test-set performance in 27 mammographers found no correlation among settings and results [18]. Our study used the same cases and test situation for the gold standard interpretations and the study radiologists' interpretations, possibly providing a more valid comparison. Our findings in a testing environment are congruent with Venkatesan, et al., who examined the positive predictive value (PPV) of specific findings in actual practice and showed that asymmetries had the lowest PPV, *i.e.*, that most recalled findings were benign [19].

Radiologists were instructed to report the most significant finding for the mammogram. In some cases, although study radiologists noted the same lesions as the expert panel as indicated by the "clicks" on the screen, they differed with the gold standard interpretation in assignment of the "most significant" lesion; i.e. the final lesion type ascribed to the case. This speculation also is consistent with the greater agreement at the woman level, i.e., although there was greater disagreement in which finding was most significant, the clinical importance assigned to the different findings led to the same action.

The images were converted from analog to digital, with some loss of image quality. The digitization process used in the study (from the American College of Radiology) is the same as was used for the Committee on Mammography Interpretive Skills Assessment (COMISA) exam (now MCR - Mammographic Case Review); however in our study, cases with findings of interest were not specifically chosen based on feature image quality. Although all participants were invited to use a study laptop, participating radiologists may have reviewed the cases on personal computers, although the software program required minimum viewing criteria.

A major strength of our study is the relatively large number of participating radiologists. In addition, the custom-made software contained important features for viewing and interpreting, including availability of comparison films, pan and zoom features, and a standard mammography image set for each exam (left and right MLO, left and right CC, and comparison images). Another key strength was the development of gold standard interpretations through a rigorous consensus process with three nationally recognized mammography experts. We had an explicit goal of creating a test set representative of clinical practice. Thus, we randomly selected exams from clinical practice, thereby including some difficult cases, which introduced more variability, but increased generalizability.

This study provides important insights into the types of mammographic cases that contribute the most to interpretation variability. By understanding the extent to which case difficulty and finding type affects interpretive agreement, we can develop targeted training modules and educational interventions that yield the greatest improvement in radiologist interpretive performance. Our analysis of mammography interpretation agreement between radiologists and an expert panel suggests that mammography training should focus on identification and correct interpretation of asymmetric densities and architectural distortion.

Acknowledgments

This work was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope® Campaign [SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274-01, SIRSG-07-275, SIRGS-06-281, SIRSG-09-270-01, SIRSG-09-271-01, SIRSG-06-290-04], the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium [U01CA63740, U01CA86076, U01CA86082, U01CA70013, U01CA69976, U01CA63731, U01CA70040, HHSN261201100031C]. The collection of cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. The authors thank Jose Cayere, Amy Buzby and the American College of Radiology for technical assistance in developing and supporting implementation of the test sets. Their work was invaluable to the success of this project. We also thank the participating women, mammography facilities and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/."

REFERENCES

1. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology. 2006; 241:55–66. [PubMed: 16990671]
2. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammography interpretation. JNCI. 2003; 95:282–90. [PubMed: 12591984]
3. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994; 331:1493–9. [PubMed: 7969300]
4. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. JNCI. 1998; 90:1801–9. [PubMed: 9839520]
5. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: Inter- and intraobserver variability in feature analysis and final assessment. Am J Roentgenol. 2000; 174:1769–77. [PubMed: 10845521]
6. Ciatto S, Houssami N, Apruzzese A, et al. Reader variability in reporting breast imaging according to BI-RADS® assessment categories [the Florence experience]. The Breast. 2006; 15:44–51. [PubMed: 16076556]
7. Antonio ALM, Crespi CM. Predictors of interobserver agreement in breast imaging using the Breast Imaging Reporting and Data System. Breast Cancer Res Treat. 2010; 120:539–46. [PubMed: 20300960]
8. Abdullah N, Mesurolle B, El-Khoury M, Kao E. Breast Imaging Reporting and Data System lexicon for interobserver agreement for assessment of breast masses. Radiology. 2009; 252:665–72. [PubMed: 19567644]
9. Lazarus E, Mainiero MB, Schepps B, et al. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology. 2006; 239:385–391. [PubMed: 16569780]
10. Venta LA, Hendrick RE, Adler YT, et al. Rates and causes of disagreement in interpretation of full-field digital mammography and film-screen mammography in a diagnostic setting. Am J Roentgenol. 2001; 176:1241–8. [PubMed: 11312188]
11. Skaane P, Engedal K, Skjennakl A. Interobserver variation in the interpretation of breast imaging comparison of mammography, ultrasonography, and both combined in the interpretation of palpable noncalcified breast masses. Acta Radiol. 1997; 38:497–502. [PubMed: 9240666]
12. Ooms EA, Zonderland HM, Eijkemans MJC, et al. Mammography: interobserver variability in breast density assessment. The Breast. 2007; 16:568–76. [PubMed: 18035541]
13. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. Am J Roentgenol. 1997; 16:1001–8. [PubMed: 9308451]
14. National Cancer Institute. [Accessed April 14, 2011] Breast Cancer Surveillance Consortium Homepage. http://breastscreening.cancer.gov/

15. Carney PA, Bogart A, Geller BM, et al. Association between time spent interpreting different lesion types and accuracy of screening mammography. Am J Roentgenol. In Press.

16. Baker JA, Kornguth PJ, Floyd CE Jr. Breast Imaging Reporting and Data System standardized mammography lexicon: observer variability in lesion description. Am J Roentgenol. 1996; 166:773–8. [PubMed: 8610547]

17. Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. 2008; 249:47–53. [PubMed: 18682584]

18. Rutter CM, Taplin S. Assessing mammographer’s accuracy: A comparison of clinical and test performance. J Clin Epi. 2000; 53:443–450.

19. Venkatesan A, Chu P, Kerlikowske K, et al. Positive predictive value of specific mammographic findings according to reader and patient variables. Radiology. 2009; 250:648–57. [PubMed: 19164116]

Take Home Points

- Interpretive agreement with an expert panel was high for cancer cases, but less so for non-cancers, particularly for subtle lesions and calcifications, architectural distortions and asymmet-ric densities.
- Asymmetric densities accounted for half of all unnecessary recalls.
- Differences in agreement were often due to nomenclature of a given lesion
- Subtle, non-mass lesions are the most challenging to interpret clinically and may benefit from targeted training strategies

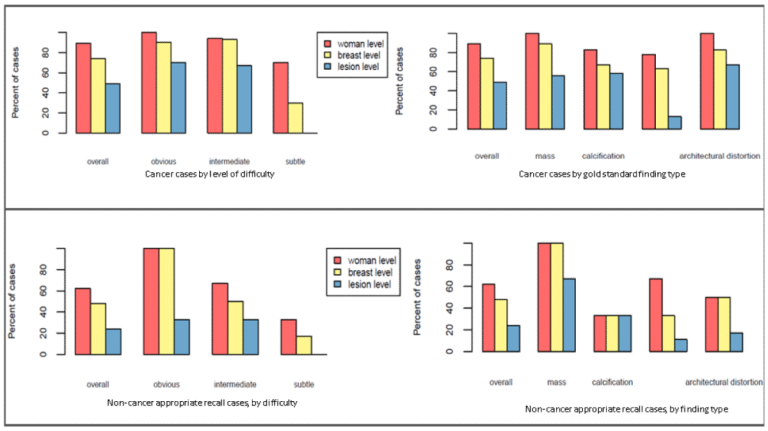

Figure 1a-d.

Percent of participating radiologists recalling cases based on level of lesion difficulty and finding type [mass, calcification, asymmetric density, and architectural distortion].

Table 1
Characteristics of participating radiologists

	Number of radiologists (%)
Total	119
Breast specialist	
Yes	9 (7.6%)
No	110 (92.4%)
Fellowship training in breast or women's imaging	
Yes	15 (12.6%)
No	104 (87.4%)
Main practice with academic radiology group	
Yes	17 (14.3%)
No	102 (85.7%)
Average days per week working in breast imaging	
1 day or less	34 (28.6%)
2 days	24 (20.2%)
3 days	28 (23.5%)
4 days	15 (12.6%)
5 days	18 (15.1%)
Years interpreting mammography	
1-5 years	25 (21.0%)
6-10 years	18 (15.1%)
11-20 years	51 (42.9%)
21-30 years	17 (14.3%)
31 years or more	8 (6.7%)
Average number of mammograms interpreted per year during the last 5 years	
Don't know	29 (24.4%)
Up to 1000	15 (12.6%)
1001 to 2500	38 (31.9%)
2501 to 4000	15 (12.6%)
4001 or more	22 (18.5%)
Mammography examinations interpreted per week	
Up to 10	8 (6.7%)
11-49	25 (21.0%)
50-99	36 (30.3%)
100-199	31 (26.0%)
200 or more	19 (16.0%)
Self-described ability to perceive and determine importance of mammographic findings	
Not sure	3 (2.5%)
Below average	2 (1.7%)

	Number of radiologists (%)
Average	53 (44.5%)
Above average	50 (42.0%)
Expert	11 (9.2%)
AIM-BCSC test set number	
1	30 (25.2%)
2	34 (28.6%)
3	28 (23.5%)
4	27 (22.7%)

Table 2

Composition of the four AIM test sets and percent of cases recalled by the community radiologists.[*,†]

	Percent recalled: median		
	Woman level	**Breast level[‡]**	**Lesion level[§]**
Overall	43	NA	NA
Cancers	80	73	67
Difficulty			
Obvious	100	100	83
Intermediate	86	86	71
Subtle	50	40	25
Expert finding type			
Mass	100	100	83
Calcification	82	70	64
Asymmetric densities	60	50	50
Architectural distortion	80	80	50
No cancer	34	NA	NA
Appropriate recalls	67	57	43
Difficulty-appropriate recalls			
Obvious	100	100	67
Intermediate	75	67	50
Subtle	33	17	17
Expert finding type-appropriate recalls			
Mass	100	100	67
Calcification	33	33	33
Asymmetric densities	56	44	33
Architectural distortion	50	50	33
Other non-cancers	26	NA	NA

[*] Radiologists labeled each case in the test set as a recall (BI-RADS assessment 0, 4, or 5: Needs additional imaging) or a non-recall (BI-RADS assessment 1 or 2: Negative or benign).

[†] The number of radiologists that interpreted each case varies across cases.

[‡] Breast level percents are the percent of cases in each category for which the radiologist recalled the same breast as did the experts (not applicable for films not recalled by the experts).

[§] Lesion level percents are the percent of cases in each category for which the radiologist recalled the same lesion/region as did the experts (not applicable for films not recalled by the experts).

Table 3

Distribution (%) of community radiologist finding types for cases in each category of expert finding type

	Community radiologist finding type			
	mass	**calcification**	**asymmetric density**	**architectural distortion**
Expert finding type				
mass	**79.9**	1.0	12.6	6.6
calcification	7.0	**91.7**	0.8	0.5
asymmetric density	40.4	4.1	**40.7**	14.7
architectural distortion	25.0	9.3	20.4	**45.3**

Table 3a

Distribution (%) of finding types: Subtle cases

	Community radiologist finding type			
	mass	**calcification**	**asymmetric density**	**architectural distortion**
Expert finding type				
mass *	NA	NA	NA	NA
calcification	0.0	**98.4**	0.0	1.6
asymmetric density	21.4	1.3	**38.4**	39.0
architectural distortion	21.6	0.0	52.9	**25.5**

* No subtle lesions were labeled a mass by the experts.

Table 3b
Distribution (%) of finding types: Intermediate cases

	Community radiologist finding type			
	mass	**calcification**	**asymmetric density**	**architectural distortion**
Expert finding type				
mass	**75.5**	1.0	15.0	8.5
calcification	0.7	**98.9**	0.0	0.4
asymmetric density	46.5	5.0	**41.5**	7.0
architectural distortion	32.1	14.6	16.1	**37.3**

Table 3c
Distribution (%) of finding types: Obvious cases

	Community radiologist finding type			
	mass	**calcification**	**asymmetric density**	**architectural distortion**
Expert finding type				
mass	**90.6**	0.9	6.6	1.9
calcification	12.4	**85.7**	1.5	0.5
asymmetric density[†]	NA	NA	NA	NA
architectural distortion	9.5	0.0	19.6	**71.0**

[†]No obvious lesions were labeled an asymmetric density by the experts.

Table 4
Distribution of finding type labels used by community radiologists on non-cancer cases

	Finding type label			
	mass	**calcification**	**asymmetric density**	**architectural distortion**
Non-cancer cases				
Appropriate recalls: recalled by experts *	0.50	0.14	0.26	0.10
Unnecessary recalls: not recalled by experts	0.27	0.07	0.51	0.15
All non-cancer cases	0.37	0.10	0.40	0.13

* Radiologists were instructed to locate and label the single most important lesion for all recalled cases. For non-cancer cases recalled by the experts, the expert finding type label may refer to a different location than that identified by the community radiologist recalling the case.